Case–Control Study

The case-control design provides a framework for studying the relationship between possible risk factors and a disease by collecting information about exposure from those with disease but only from a fraction of the individuals under study who do not develop disease. When the disease is rare, this approach offers a major gain in efficiency relative to the full cohort study, in which an investigator seeks information on exposure for everyone. The savings compensate handsomely for the loss in the precision of estimates of parameters describing the relationship between exposure and disease that could have been obtained from studying everyone. In fact,

Wacholder S, Hartge P. Case-control studies. In Armitage P, Colton T, eds. Encyclopedia of biostatistics. New York: Wiley, 1998;503-14.

Table 1 Hypothetical full cohort study

	Diseased	Nondiseased	Total	Relative risk[a]	Relative odds[b]
Exposed	16	99 984	100 000	3.60	3.60
Unexposed	40	899 960	900 000	1.00	1.00

[a] Relative risk = (16/100 000)/(40/900 000).
[b] Relative odds = (16/99 984)/(40/899 960).

the reduction in precision often is marginal. By collecting data on exposure about *cases*, the subjects who have developed disease, and **controls**, specially selected subjects without disease, the case-control design also compresses the time needed to complete the study. In a classic case-control study, Doll & Hill [10] recruited 649 male lung cancer cases and 649 male controls during an 18-month period in London. They were able to show a clear increase in **risk** with increasing daily cigarette consumption in this case-control study (*see* **Smoking and Health**). By contrast, in a cohort study of an equal number of men at the very highest risk - that is, very heavy smokers above age 70 - one would expect to find only a handful of lung cancer cases within 1.5 years, not nearly enough to draw convincing conclusions about the relationship between smoking and lung cancer.

A hypothetic example illustrates the extent of the savings. In Table 1 are displayed the results of a cohort study of 1 000 000 individuals who are followed for disease for one year; 10% of them are exposed.

The expected results from a case-control study in which all 56 of the cases from this cohort are studied are displayed in Table 2. Expected cell counts are also shown in Table 2. For example, the expected number exposed among the 56 studied controls is calculated as 56 × (99 984/999 944) = 5.6.

The estimate of the **odds ratio** for disease, (16/40)/(99 984/899 960) in Table 1, equals the estimate of the exposure odds ratio in Table 2, (16/40)/(5.6/50.4). Both odds ratios equal 3.6004, and approximate the risk ratio (*see* **Relative Risk**) (16/100 000)/(40/900 000) = 3.60 to four significant digits. Thus, the study of 112 individuals would give the same estimate as the study of 1 000 000, apart from random variation. While the 95% **confidence interval** (CI) for the odds ratio from the case-control study, (1.3-10.3), is substantially wider than the CI (2.0-6.4) from the full cohort study, using 5 × 56 = 280 controls instead of only 56 would narrow the CI for the case-control study to (1.8-7.2), which is notably closer to that of the full cohort study. This minor loss of precision is a small price to pay for the savings in exposure assessment costs and in time that may make feasible a study that would otherwise be too expensive.

Table 2 Expected values from case-control study in same setting as Table 1

	Diseased (cases)	Nondiseased (controls)	Relative odds
Exposed	16	5.6	3.60
Unexposed	40	50.4	1.00
Total	56	56	

In principle, although not always in practice, all case-control studies yield an **unbiased** estimate of the odds ratio and other functions of the odds. Most are designed so that the odds ratio directly estimates the relative risk or the incidence-rate ratio. However, only **population-based case-control studies** that yield estimates of overall disease risk or rate in the population permit estimation of exposure-specific **incidence rates** and thus of all parameters that could be estimated from studying the entire cohort.

Along with these considerable design strengths, the case-control study has several weaknesses. Incomplete or inaccurate ascertainment of outcome and improper selection of controls can cause **selection bias**. Retrospective assessment of exposure history can lead to **nondifferential** and **differential** measurement error and **biased** estimates of exposure effects. As in any nonexperimental or **observational study**, **confounding** can distort the estimates of effect from a case-control study (*see* **Bias in Case-Control Studies; Bias in Observational Studies; Bias, Overview; Measurement Error in Epidemiologic Studies; Misclassification Error**).

The Range of Case-Control Studies

A MEDLINE search for papers published since 1992 found over 1500 entries per year mentioning case-control or one of its cognates, usually case-referent. The case-control study is a fundamental tool of epidemiology with broad application in areas as diverse as the etiology of cancer and birth defects, the effectiveness of vaccination and **screening** for disease, and the causes of automobile accidents.

Case-control studies vary greatly in scope, sources of data, and complexity. At one extreme are investigations of an outbreak, which may include fewer than ten cases (*see* **Communicable Diseases**). These studies often encompass a wide-ranging, open-ended examination of many exposures and host characteristics of the cases. Often, the selection of controls can precisely correspond to the source of cases because there is a roster of the source population (for example, in a hospital outbreak) or a convenient collection of willing participants. At the other extreme are multicenter, multiyear, highly focused studies of tens of thousands of cases and controls. These are not common, because of their high cost. More typical are studies of a few hundred cases and an equal number of controls selected without a roster, but with an **algorithm** intended to represent the population from which the cases arose. These intermediate-sized studies provide a practical approach when the relative risk is expected to be around 2 or greater and the exposure is reasonably common (10% or more).

Weaknesses of the Case-Control Approach

Case-control studies, like **cross-sectional** and observational cohort studies, suffer from the common drawbacks of all nonexperimental, or observational, research, stemming from the investigator's lack of control in assigning exposure. Foremost is the absence of **randomization** as a tool for reducing confounding. An observational study will not be as reliable as a **clinical trial** for investigating questions such as the effectiveness of a new treatment or screening program.

Even though a case-control study has no intrinsic shortcomings compared exposuto a nonexperimental full cohort study that collects information on everyone in the same setting, the case-control design has often been disparaged as fundamentally weaker than the full cohort study. Several conceptual, statistical, and practical reasons explain this negative attitude. Many early observers saw the case-control study as a "backward cohort study", with inference made from effect to cause. It was not obvious how to translate a difference in exposure between cases and controls into a parameter describing prospective risk until **Cornfield** in 1951 [7] showed theoretically that the exposure odds ratio from a case-control study approximates the disease risk ratio from a case-control study when the outcome is rare. Selection bias can arise from poor study design or poor implementation in choosing cases and controls. Retrospective ascertainment of information about exposure and confounders may yield inaccurate data leading to bias. These issues are discussed in detail later in this article.

Another apparent weakness of the case-control approach is that ordinarily it yields relative but not absolute measures of the effect of exposure on disease. It is possible, however, to estimate exposure-specific **absolute risk** and risk differences when the crude risk of disease is known in the study population [1, 7, 10, 30].

Case-Control Study as a Missing-Data Problem

A population-based case-control study can be regarded as a cohort study with many nondiseased subjects missing at random [30]. This view of the case-control study helps to resolve many conceptual issues. It reveals when and how a broader class of parameters, including absolute risk and risk difference, can be estimated. It clarifies the requirements for proper control selection (*see* **Missing Data in Epidemiologic Studies; Missing Data**).

Consider a population-based case-control study to examine the effect of an exposure on the risk of developing disease. In the ideal study, the investigator is able to identify all cohort members newly diagnosed with disease during a specified follow-up period. These people with disease, or a random subset, become the cases in the study. Controls are a **random sample** of the noncases. The investigators obtain information on exposure that preceded the time of onset of disease from these cases and controls. Exposure information for those noncases who never develop disease during the study period will be *missing at random* if the investigator determines whose exposure will be collected, based only on disease status, which is known for individuals during the specified time. Thus, the case-control study is a missing-data problem, albeit with two unusual features: the "missingness" is a planned maneuver rather than an uncontrollable accident, and the ratio of missing to observed data can be extraordinarily high.

Under these assumptions, the cases and controls will have the same exposure distribution as the diseased and nondiseased, respectively, in the cohort, and the investigator can estimate from the case-control data all of the parameters estimable from the full cohort study. Indeed, under these assumptions, there are no intrinsic weaknesses to the case-control design. This outlook recognizes the prospective nature of the study, allows estimation of all parameters available from the full cohort, including absolute risk and risk difference, and demonstrates why the controls selected should have the same exposure distribution as other nondiseased individuals in the study population [30]. The inference from the missing data approach is identical to standard case-control inference in this setting [30].

Case-Control Studies to Estimate a Hazard Ratio

In the idealization described above, risk is described as the probability of developing disease during a fixed interval. If the study aims to estimate functions of **hazard rates** of disease, or numbers of new events per unit of person-time (*see* **Person-Years At Risk**), the time element must be incorporated more precisely. For instance, in the standard **proportional hazards** analysis of the full cohort study designed to estimate the hazard ratio, the **partial likelihood** compares the exposure of a case to that of the members of the *risk set*; namely, all other members of the cohort who are at risk at the time of the event that defines when the cohort member became a case.

In the **nested case-control** study that would be undertaken in the same cohort, as first described by Thomas [27], exposure from only a few randomly selected members of each risk set is collected and used in a time-matched case-control analysis, an analog to partial likelihood. Again, except for the use of fewer individuals, there is no intrinsic difference between the full cohort and nested case-control analyses. All noncases in the risk set should be eligible and equally likely to be sampled as controls, even those who were previously selected as controls or who later develop disease [15]. Sampling at event times should be mutually independent in the nested study.

The **case-cohort design**, first described by Prentice [21], is a useful alternative with several practical advantages. The controls are selected as a single sample or *subcohort* from the entire cohort, including cases. While the sampling is not time-matched, in the analysis the **likelihood** at each event time uses the exposures of the case and of the subcohort members who are in the risk set at the event time. The fact that the subcohort is a random sample of the cohort leads to more flexibility in the analysis and allows the same controls to be used for analyses of several endpoints.

Design

There are three interlocking steps in planning the design of a case-control study:

1. Investigators must decide whether a cohort or case-control study is appropriate.
2. Investigators must determine who will be cases and controls in the study and how to assess exposure.
3. Investigators must decide on all the specific details to be included in the study protocol.

Full Cohort vs. Case-Control?

The first decision required in planning a case-control study is to determine whether the case-control design is more appropriate than a full cohort design [29]. The reasons for preferring a case-control study to the full cohort study are almost always practical, revolving around feasibility, economy, speed, and the need to study multiple exposures or their joint effects. On the other hand, a prospective cohort study sometimes affords an opportunity to collect more reliable exposure information, and can be used to study **multiple health outcomes** simultaneously. It can offer slightly more statistical precision. Finally, justifiably or not, the cohort study has more credibility.

Lower Cost vs. Higher Statistical Efficiency. Studying fewer subjects reduces the cost but also lowers the precision of the estimate of effect. When the disease is rare, the impact will be very modest, as the above example demonstrates. The **variance** estimate of the log-odds ratio estimate from **two-by-two tables** of the form in Table 1 or Table 2 is the sum of the reciprocals of the cell entries, so the size of the smallest cell in the two-by-two table is the factor limiting precision. When exposure is rare, this smallest cell almost always will be the number of exposed cases. This quantity is the same in the full cohort or in the case-control study performed in the same setting. Thus, the relative efficiency of the case-control study with k controls per case is $k/(k+1)$ compared to the full cohort [28]. The choice of design often boils down to whether to look for cases among the exposed (as in a cohort study) or exposed among cases (as in the case-control approach).

The clearest advantage for the case-control study occurs when the outcome of interest is rare and the exposure of interest is common. As the percentage of individuals experiencing the outcome during the follow-up period increases, the efficiency advantage of the case-control design diminishes. As the exposure of interest becomes rare, the ability of the case-control study to estimate an effect diminishes and a cohort design that ensures that individuals with the rare exposure will be followed for disease may become more advantageous.

Data Quality. Exposure assessment is the Achilles heel of the case-control study. If information collected retrospectively about exposure is of lower quality than concurrent data, more *nondifferential* **misclassification** or error, and consequently, attenuation of estimates of effect, almost inevitably ensue. Worse still, exposure information that is self-reported is susceptible to *differential* error or misclassification, namely different error patterns in cases and controls.

The resulting bias can work to exaggerate, attenuate, or reverse the direction of an effect. While differential error from interviews has been difficult to establish conclusively in particular situations, it seems realistic to assume that the accuracy and thoroughness of reports from cases, who are touched by the research question and whose lifestyle may be affected by the disease, will be greater than for controls. The effect of differential error is often called report or **recall bias**. Some nutritional epidemiologists are extremely skeptical of dietary data collected from cases and controls retrospectively, for fear of differential misclassification (*see* **Nutritional Exposure Measures**). By contrast, when previously written records are the source of exposure information, the errors are no different from those in a full cohort study. So a retrospective or even a prospective full cohort study would not automatically have higher data quality. Correspondingly, collection of reliable information on outcomes in all members of a cohort or a case-control study is also a challenge, especially for softer endpoints, such as infertility.

Other Scientific Issues. Apart from considerations of efficiency, reflecting the rarity of disease and exposure, other considerations come into play. When confounding poses a major problem for a study, accurate confounder assessment may dictate one design or the other. The need to study multiple exposures magnifies the advantage of the case-control design, while a cohort study allows additional outcomes to be included in the study with little increase in cost. Some well-established cohorts [37] have demonstrated that results on the relationships between multiple exposures and multiple exposures from a full cohort study can justify its substantially greater cost relative to a single case-control study.

Credibility. While most researchers and journals now appreciate the case-control design, some still consider case-control studies automatically suspect [11]. While this attitude is becoming less widespread, it may affect how one's work is accepted.

Choice of Setting

The specific setting for the study must be chosen within constraints imposed by logistics, convenience, and cost. Investigators must also consider the key factors that determine the quality of a case-control study in a particular location. How complete and accurate will the case ascertainment be? How rapidly will investigators receive reports of cases, thereby reducing the influence of the postdiagnosis period, such as effects of treatment, and the number of fatal or debilitated cases who might be excluded or whose exposure information may need to be collected from a proxy, such as a spouse or child? Is there a roster or **sampling frame**, possibly from electoral lists or a health insurance plan (*see* **Administrative Databases**), from which to select suitable controls? Are written records available to evaluate exposure, thereby reducing the possibility of differential misclassification? Are participants likely to give reliable information on exposure or confounders, including perhaps family medical history, prescription drug use, or highly personal questions about sexual history or a previous abortion? Are participation rates likely to be high? (*see* **Nonresponse**). Will participants be amenable to a procedure needed for the study, such as blood drawing for assessing a biomarker? What is the rate of occurrence of events and how will it affect the amount of time needed in the field? Is there enough heterogeneity of exposure to reduce the cost of a study and the number of subjects needed to achieve a specified precision?

Case-control studies can be oriented toward measuring the effect of exposure on disease **prevalence**, **cumulative hazard**, or **incidence rate**. Thus, the temporal perspective must be considered. Ought the study be limited to future cases, or can previously diagnosed individuals be used? Using only those cases that are newly diagnosed (**incident cases**) generally works to improve case ascertainment and participation, reduce reliance on proxy respondents for deceased or disabled cases and simplify control selection, but is slower and more costly. One subtly different definition of cases produces an estimate of cumulative risk rather than **incidence density ratio**; namely, when cases are all subjects who developed disease throughout the duration of follow-up of a population. Finally, diseases with poorly defined onset and long duration call for prevalence studies, with the definition of cases correspondingly changed to subjects who have the disease at the specified point in time, regardless of when they first developed it (*see* **Case-Control Study, Prevalent**).

Case and Control Selection

Case and control selection must be defined together because they are intrinsically linked. Miettinen's [20] concept of the *study base* helps to clarify this connection. The study base at a given time consists of those individuals who would become cases in the study if they developed disease at that time. When the study base is well-defined, the study is called a *primary-base* study or a *population-based case-control* study; cases are simply those members of the study base who experience the outcome and controls can be a random sample from the base. In this situation, it is possible to determine whether any individual is in or out of the study base at a given time and whether that individual is eligible to be a case or control in the study. The problem is making sure that all cases in the base come to the attention of the study investigators. The alternative starts with a set of cases, perhaps chosen for convenience, as in a *hospital-based case-control* study of lung cancer diagnosed at a single hospital during a single year. In these *secondary-base* studies, the study base is poorly defined because it is not always clear whether an individual who did not develop disease would have been a case in the hypothetic circumstance of development of disease. With no way to know whether a potential control would have come to the study hospital upon development of disease, random sampling for control selection is impossible. Thus, these secondary-base controls must be *assumed* to be an approximation to a hypothetic random sample that could characterize the study base. So in the primary base study, the difficulty is finding the cases, while, in the secondary base study, the difficulty is ensuring an appropriate set of controls.

Case Selection. In the idealized case-control study, all subjects with disease in the study base (or a random sample of them) become cases. In reality, some cases do not come to the attention of the investigators, some individuals are falsely called cases when in fact they do not meet the diagnostic criteria, and some eligible cases refuse to participate. In a study of male infertility, factors that lead to someone to regard lack of children as a problem might appear as risk factors because of differential case ascertainment [31]. Inaccurate and incomplete case ascertainment can create selection bias as well as reduce precision. When there is ambiguity as to whether someone truly developed disease, as in the absence of a definitive pathology report, the standard practice of excluding the case is not harmless, if those lacking information have different exposure distributions than those with the information, perhaps because they are seen at a hospital in a poorer area [29].

Principles of Control Selection. There are three principles that underlie control selection: *study-base*, *comparable-accuracy*, and *deconfounding* [31]. The essence of the study-base principle is that controls can be used to characterize the distribution of exposure in the study base from which the cases arise. The comparable-accuracy principle calls for equal reliability in the information obtained from cases and controls so that there is no *differential* misclassification. Thus, a study of drug use during pregnancy as a risk factor for a specific type of birth defect might call for a control group of children who experienced a comparably serious outcome at birth so that the mothers of cases and controls would be equally likely to recall exposure during pregnancy accurately. The deconfounding principle allows elimination of confounding through control selection, such as through matching or **stratified sampling**, to be a consideration in control selection. These principles may conflict with one another and may have strong negative impacts on efficiency. They should not be regarded as absolute, but rather as points to consider in choosing a control group.

Controls for Studies with a Roster. In fortuitous situations, the investigator can use a roster listing all individuals and the period when they are in the study base. Investigators can then sample at random from the roster to satisfy the study-base criterion.

Controls for Primary-base Studies without a Roster. When a roster is not available and cannot be created from electoral or town residence lists, it is impossible to generate a random sample directly. A commonly used approach when there is no roster is **random digit dialing** (RDD) [34], an efficient way to generate a near-random sample often used in public opinion polling. RDD relies on dialing telephone numbers according to a strategy that yields representative samples. RDD suffers from several potential biases. RDD will not select individuals without phones, although it can compensate for households with multiple telephone lines. Furthermore, many people refuse to respond to telephone surveys, especially since the advent of answering machines (*see* **Telephone Sampling**). Empirically, controls chosen by RDD seem to be of higher socioeconomic class than a truly random sample would be. This violation of the study-base principle may be alleviated by adjustment for income or socioeconomic status.

Requirements for individual controls vary. In *incidence-density sampling* [14, 19, 22], used most commonly in primary-base studies, controls must be disease free at the time of diagnosis of the case to which they are matched. As in the nested case-control study, this design allows estimation of an incidence rate-ratio (and **relative hazard**) and eliminates the need for the rare-disease assumption [14, 19]. For *cumulative-incidence sampling*, controls are selected from among those who survive the study period without developing disease. Cumulative-incidence sampling of controls allows estimation of the risk ratio (**relative risk**), which approximates the relative hazard only when the rate of disease is low.

Secondary-Base Studies. Some diseases, including those not consistently detected in the general population, dictate an alternative to primary-base studies. For example, when case identification is incomplete, population controls may not be appropriate when completeness of case identification is differential by exposure and the selection bias cannot be corrected by adjustment for another variable. The most common secondary-base study is the *hospital-based case-control* study. Controls are patients seen at the same hospital as the cases, but for a different condition. This approach works well when two requirements are met. First, both cases and controls must be people who would have presented at the same hospital if they had *either* the case-defining illness or the control-defining condition. Secondly, the conditions used to select controls cannot be associated with the exposure. If these requirements are met, the distribution of the exposure in the controls reflects the distribution in the study base. The investigator seldom knows with certainty that both criteria are met, so compliance with the study-base criterion remains hard to verify convincingly.

A possible advantage of the hospital-based control group is more confidence that the equal accuracy criterion will be met. With equally serious illnesses, cases and controls ought to provide similarly complete and accurate reporting of past exposures. Thus, for the study of a specific birth defect, controls could be chosen from babies born with another birth defect of similar severity but known not to be related to the exposure of interest. Using controls with cancer at other sites for a study of a form of cancer may help with the equal-accuracy principle, but care must be taken so that cancer at the control site is not related to exposure.

Other Kinds of Control Groups. While population and hospital controls are the most commonly used kinds of control groups, investigators have used other options [32]. Use of patients from the *same primary care provider* as the case helps to insure that a control who developed the disease of interest would have become a case in the study. Use of *friends* of the cases can lead to bias in studies of factors related to sociability. Use of *relatives*, often siblings, as controls may reduce confounding by genetic factors. Each of these control groups requires a careful selection procedure to make sure that individuals are not being picked to be controls in a way that is related, directly or indirectly, to the factors under study.

Design Options. **Matching** on well-established confounders is a common practice in case-control studies. In case-control studies, matching serves to increase the precision of the estimated effect of exposure by making the distribution of the confounder identical in the cases and controls. Usually, the efficiency advantage from matching is small, and may not compensate for the extra cost and complexity, the exclusion of cases for whom no match is found, and the reduced flexibility of the analysis [33]. Other justifications of matching include control for nonquantitative variables such as neighborhood and the ability to control for confounding without making assumptions about the effect of the confounder in the risk model [33].

Only strong confounders should be considered as matching variables. Two-phase designs, discussed below, are more appropriate if one wants to estimate the effect of a variable considered for matching. Demographic variables such as race and sex and temporal variables such as age and calendar year (or decade) of first employment are the most suitable matching variables. Matching is always inappropriate on a factor that is a consequence of exposure.

Two-Phase Designs. These techniques [2, 36] (*see* **Case-Control Study, Two-Phase**) are a more flexible generalization of matching, also used to increase efficiency or to reduce the cost of exposure assessment. In two-phase designs, detailed information on exposures and confounders is not ascertained for everyone, but only for subsets of cases and controls, with the selection probability depending on case status and on the value of another variable that is available for everyone. Instead of requiring, as in matching, that the distribution of the variable be the same in the control as in the cases, essentially arbitrary distributions in each group are specified. These two-stage designs allow the estimation of both main effects and **interactions**. For example, in a study designed to investigate the joint effects of domestic radon exposure, requiring expensive measurements, and smoking, which is easier to ascertain, on the risk of lung cancer, taking all cases and a random sample of controls would lead to a study with a preponderance of smoking cases and nonsmoking controls; matching controls to cases on smoking status would lead to small numbers of control nonsmokers as well. The assessment of interaction is much more efficient in the two-phase design where nonsmoking cases and smoking controls are oversampled [35].

Sample Size. There is an extensive literature on **sample size determination** for case-control studies [4, 24, 25]. As in the full cohort study, needed sample size is dependent on the variation in exposure in the study base. A key point is that increasing the ratio of controls to the harder-to-find cases increases the precision of the odds ratio estimate in an increasingly marginal way, especially for small effects. Ratios of controls to cases beyond four or five are usually not advisable because the successive gains in efficiency diminish. Indeed, the **asymptotic relative efficiency** for a study involving k controls per case is $k/(k+1)$, which takes on values of 0.5, 0.67, 0.75, 0.8, and 0.83 for k from 1 through 5 [28].

Fieldwork

The best-designed study will not be convincing unless the fieldwork is sound. In the field, case-control studies face the usual challenges of observational research: identifying all members of the study population, achieving an adequate response rate, collecting accurate data, and measuring potential confounders.

Most case-control studies include a questionnaire, because seldom have all of the exposure variables of interest been recorded in documents easily available to the investigator. Sometimes the study subject, or his surrogate, completes the questionnaire ("self-administered"); alternatively, an interviewer can pose the questions. A questionnaire can be computerized or on paper; an interview can be in-person or by telephone (*see* **Computer-Assisted Interviewing; Interviewing Techniques; Questionnaire Design**). Depending on the hypotheses, investigators may also collect biologic specimens, samples of the study subject's present or past environment, and permission to contact agencies that have documented data about the exposures.

The case-control design poses some specific problems, as well. Since the cases have already developed the disease, it will not be possible to estimate the effects of exposure measures that are distorted by the disease, including weight and body biochemistry, unless the investigator has access to stored measures that were collected before disease onset. If it is not clear whether a measure is likely to be valid once the disease is clinically manifest, the investigator may conduct a specific methodologic pilot study. Sometimes, it is possible to examine the effects specific for stage of disease, in the expectation that post-onset distortions will be more pronounced with more advanced disease. In a similar fashion, the investigator will consider whether therapy influences the level of the exposure variable. If so, then cases need to be studied before therapy begins, or well after any of its influence has waned.

Just as diagnosis and treatment of a serious disease can cause biological changes in exposure variables, they also can cause changes in a patient's recollection or willingness to report various exposures. The resulting **recall bias** does not always go in a particular direction; the specific exposure needs to be considered, preferably with data on reporting bias from ancillary sources. Some exposures lend themselves to internal validation by studying a higher-quality exposure variable on a subset of subjects or by collection of validation data from other sources, such as medical records (*see* **Validation Study**). In that circumstance, some or all of the subjects reporting an illness or hospitalization will be asked to give permission for review of records; ideally, some of the reports of no hospitalization ought to be selected for review, too, although this is seldom practical. To minimize recall bias, the investigator also attends to the exact phrasing of questions, trying to leave very little room for interpretation or rumination. Sometimes investigators attempt to blind the interviewer to the case-control status of subject, but often the status of the subject becomes apparent anyway (*see* **Blinding or Masking**).

With access to prospectively collected data stored in records, the investigator can avoid the problem of differential misclassification stemming from the fact of diagnosis. Even with stored records, however, one source of differential misclassification could be present: minor abnormalities noted because of greater medical surveillance of the exposed may not have been detected in the unexposed (*see* **Bias from Diagnostic Suspicion in Case-Control Studies; Bias from Exposure Suspicion in Case-Control Studies**).

Analysis

The goal of the analysis of case-control studies is almost always to identify risk factors that are related to disease and to determine whether in fact the risk factors are causes of the disease (*see* **Causation**). As in other nonrandomized situations, the analysis must address the possibility of *confounding* and **effect-modification** by **measured covariates**.

The primary difficulty that is inherent to analysis of case-control studies is that the sampling is based on disease status while the parameters of interest relate to risk or rate of disease. Thus, it is not the difference in exposure frequency or **means** between cases and controls that is of direct interest, but estimates of the effect of determinants of disease on the rate of disease or on the probability of developing disease.

The analysis of case-control data can be exquisitely simple or tremendously complex. When the exposure and disease are each dichotomous (*see* **Binary Data**) and there are no other factors to consider, the analysis reduces to a **two-by-two table** of exposure by disease status. Originally, Cornfield [7] proposed that the *odds ratio*, or cross product ratio in the two-by-two table could be used as an estimate of the risk ratio (or relative risk) when the disease was rare. **Mantel & Haenszel** [16] developed an estimator and a test statistic that could be used when combining tables over several strata, thereby controlling for confounding. Exact conditional approaches, not relying on asymptotic theory, are also available for obtaining inference on the common odds ratio, adjusted for confounders by

stratification [3, 13, 18] (*see* **Exact Inference for Categorical Data**).

While **discriminant analysis** seems a natural tool to distinguish cases and controls, **logistic regression**, in which the dependent variable is the logarithm of the odds of disease, has two distinct advantages. It allows for exposures, confounders and effect-modifiers that are discrete or continuous, regardless of distribution [9] and yields valid estimates of relative-odds parameters from case-control data. Prentice & Pyke [23] proved that prospective logistic modeling - that is, of disease as a function of exposure - estimated relative risk parameters correctly and with full efficiency. If the sampling fractions of cases and controls are known, as in some population-based case-control studies, the intercept estimate from the case-control analysis can be combined with the ratio of the sampling fractions to yield a valid estimate of absolute risk. Logistic regression is now the most commonly used approach to analyze case-control data. Carroll et al. [6] extended the Prentice-Pyke result to show that many variations of case-control designs could be analyzed by logistic regression and given a prospective interpretation. Extensions to the logistic framework allow the handling of more complex sampling schemes, such as two-phase designs, of **nonlinear regression** effects of covariates, and of alternative models of joint effects of two risk factors, such as **additive** rather than **multiplicative** effects.

Control for a small number of categorical confounders can be achieved by the Mantel-Haenszel estimator of the odds ratio and corresponding **hypothesis test**. These simple procedures have excellent statistical properties. Nonetheless, logistic modeling is used routinely, because of its greater flexibility, for instance, in handling continuous variables [3]. In most modern studies, there will be more than two levels of exposure or one or more confounders and effect-modifiers to consider.

The analysis of matched pairs with a single dichotomous exposure variable uses only discordant pairs. It takes the ratio of pairs with the case exposed to those with the control exposed as the odds ratio estimate. The corresponding test of the **null hypothesis** that the odds ratio is one is equivalent to the hypothesis that the number of pairs with exposed cases among the discordant ones is **binomial** with probability 0.5 (*see* **Matched Analysis; McNemar Test**). While several extensions to more complex exposure variables and matching schemes were developed, the breakthrough in the analysis of matched data was the introduction by Breslow et al. [5] of **conditional logistic regression**, which allows general matching schemes, arbitrary exposures, continuous or discrete confounders (other than those used in the matching), and effect-modifiers.

An important variation of the analysis of case-control data allows for estimation of a hazard ratio rather than an odds ratio or risk ratio, as in nested case-control and case-cohort studies [21, 26, 27]. These designs are particularly useful when exposures vary with time, as does, for example, lifetime exposure to an environmental or occupational chemical. A conditional **likelihood** approach can be used here as well as where the matching is on time. In the contribution to the conditional likelihood at each event time, the exposure is accumulated only until the event time, exactly as if the analysis were prospective and no future data were available [21]. Furthermore, the same structure of the conditional likelihood as in the matched case-control study is used. As long as the controls are selected randomly from those at risk - that is, including future cases and independently of past use as a control or time of follow-up - the estimates of hazard ratio are valid, again reflecting the close relationship to the full cohort design. When the same controls are used for each case diagnosed during the control's follow-up, as in the case-cohort design, the estimates of the hazard ratio are also valid, but the variance estimate is more complex because the scores at each event time are not independent.

In most reports of case-control studies, no estimate of *absolute risk* or *absolute rate* is given. Methods are available for population-based case-control studies when the crude risk of disease is known [1, 7], and generally for nested case-control and case-cohort studies [17, 37]. Furthermore, risk or rate difference and other nonlogistic models can be fit [30]. Methods for estimating the attributable risk and its variance in a general setting are also available [8].

Summary

The case-control study remains the most popular approach in **analytic epidemiology** because of its relatively low cost and high speed. Ascertainment of disease, selection of controls and measurement of exposure present substantial difficulties in almost every case-control study, but a large body of epidemiologic theory and experience provides guidance to meet these challenges.

References

[1] Benichou, J. & Wacholder, S. (1994). Epidemiologic methods: a comparison of the approaches to estimate exposure-specific incidence rates from population-based case-control data, *Statistics in Medicine* **13**, 651-661.

[2] Breslow, N.E. & Cain, K.C. (1988). Logistic regression for two-stage case-control data, *Biometrika* **75**, 11-20.

[3] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research.* Vol. I: *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

[4] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. II: *The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer, Lyon.

[5] Breslow, N.E., Day, N.E., Halvorsen, K.T., Prentice, R.I. & Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies, *American Journal of Epidemiology* **108**, 299-307.

[6] Carroll, R.J., Wang, S. & Wang, C.Y. (1995). Prospective analysis of logistic case-control studies, *Journal of the American Statistical Association* **90**, 157-169.

[7] Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix, *Journal of the National Cancer Institute* **11**, 1269-1275.

[8] Coughlin, S.S., Benichou, J. & Weed, D.L. (1994). Attributable risk estimation in case-control studies, *Epidemiological Reviews* **16**, 51-64.

[9] Cox, D.R. (1966). Some procedures connected with the logistic qualitative response curve, in *Research Papers in Statistics: Festschrift for Neyman J.*, F.N. David, ed. Wiley, New York, pp. 55-71.

[10] Doll, R. & Hill, A.B. (1950). Smoking and carcinoma of the lung: preliminary report, *British Medical Journal* **2**, 739-748.

[11] Feinstein, A.R. (1988). Scientific standards in epidemiologic studies of the menace of daily life, *Science* **242**, 1257-1263.

[12] Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C. & Mulvihill, J.J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually, *Journal of the National Cancer Institute* **81**, 1879-1886.

[13] Gart, J.M. (1970). Point and interval estimation of the common odds ratio in the combination of 2 × 2 tables with fixed marginals, *Biometrika* **57**, 471-475.

[14] Greenland, S. & Thomas, D.C. (1982). On the need for the rare disease assumption in case-control studies, *American Journal of Epidemiology* **116**, 547-553.

[15] Lubin, J. & Gail, M.H. (1984). Biased selection of controls for case-control analyses of cohort studies, *Biometrics* **40**, 63-75.

[16] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719-748.

[17] McMahon, B. (1962). Prenatal X-ray exposure and childhood cancer, *Journal of the National Cancer Institute* **28**, 1173-1191.

[18] Mehta, C.R., Patel, N.R. & Gray, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2 × 2 contingency tables, *Journal of the American Statistical Association* **80**, 969-973.

[19] Miettinen, O.S. (1976). Estimability and estimation in case-referent studies, *American Journal of Epidemiology* **103**, 226-235.

[20] Miettinen, O.S. (1985). The "case-control" study: valid selection of subjects, *Journal of Chronic Diseases* **38**, 543-548.

[21] Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1-11.

[22] Prentice, R.L. & Breslow, N.E. (1978). Retrospective studies and failure time models, *Biometrika* **65**, 153-158.

[23] Prentice, R.L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies, *Biometrika* **66**, 403-411.

[24] Schlesselman, J.J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York.

[25] Self, S.G. & Mauritsen, R.H. (1988). Power/sample size calculations for generalized linear models, *Biometrics* **44**, 79-86.

[26] Sheehe, R.R. (1962). Dynamic risk analysis in retrospective matched pair studies of disease, *Biometrics* **18**, 323-341.

[27] Thomas, D.C. (1977). Addendum to a paper by Liddell, F.D.K., McDonald, J.C. & Thomas, D.C., *Journal of the Royal Statistical Society, Series A* **140**, 483-485.

[28] Ury, H.K. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data, *Biometrics* **31**, 643-649.

[29] Wacholder, S. (1995). Design issues in case-control studies, *Statistical Methods in Medical Research* **4**, 293-309.

[30] Wacholder, S. (1996). The case-control study as data missing by design: estimating risk differences, *Epidemiology* **7**, 144-150.

[31] Wacholder, S., McLaughlin, J.K., Silverman, D.T. & Mandel, J.S. (1992). Selection of controls in case-control studies, I. Principles, *American Journal of Epidemiology* **135**, 1019-1028.

[32] Wacholder, S., Silverman, D.T., McLaughlin, J.K. & Mandel, J.S. (1992). Selection of controls in case-control studies, II. Types of controls, *American Journal of Epidemiology* **135**, 1029-1041.

[33] Wacholder, S., Silverman, D.T., McLaughlin, J.K. & Mandel, J.S. (1992). Selection of controls in case-control studies, III. Design options, *American Journal of Epidemiology* **135**, 1042-1051.

[34] Waksberg, J. (1978). Sampling methods for random digit dialing, *Journal of the American Statistical Association* **73**, 40-46.

[35] Weinberg, C.R. & Sandler, D.P. (1991). Randomized recruitment in case-control studies, *American Journal of Epidemiology* **134**, 421-432.

[36] White, J.E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology* **115**, 119-128.

[37] Willett, W.C., Stampfer, M.J., Colditz, G.A., Rosner, B.A. & Speizer, F.E. (1992). Relation of meat, fat and fiber intake to the risk of colon cancer in a prospective study among women, *New England Journal of Medicine* **323**, 73-77.

SHOLOM WACHOLDER & PATRICIA HARTGE